8. The NCBI BookShelf: Searchable Biomedical Books

by Bart Trawick, Jeff Beck, and Jo McEntyre

Summary

The BookShelf is a collection of biomedical books that can be searched directly in Entrez or found via keyword links in PubMed abstracts. Books have been added to the BookShelf in collaboration with authors and publishers, and the complete content (including all figures and tables) is free to use for anyone with an Internet connection.

The online books are displayed one section at a time, with navigation provided to other parts of the current chapter or to other chapters within the book. Many of the books on the BookShelf can be browsed without any restriction at all; others have less flexibility for navigating the complete content. The publisher (or the owner of the content) defines the rules for access.

The books are linked to PubMed through research papers citations within the text. In the future, more links may be established between the BookShelf and other resources at NCBI, such as gene and protein sequences, genomes, and macromolecular structures.

Content Acquisition

Basis for Inclusion

The BookShelf provides a venue through which publishers and authors can make the full text of biomedical books available to the scientific community. As the BookShelf grows, we welcome proposals from authors, editors, or publishers of any "in-scope" texts, from undergraduate textbooks to more specialized publications, including collections of review articles or workshop proceedings. The scope of the BookShelf is broadly biomedical, including clinical works and those concerning basic biological and chemical sciences.

Information for Authors, Editors, and Publishers

All books made available on the BookShelf have been provided in electronic format to NCBI. The publisher must provide the content *in full*. Because the complete content is required for display and indexing purposes, the authors, editors, and publisher of a book should all be a part of the decision to participate in the BookShelf. Interested parties can contact us at books@ncbi.nlm.nih.gov for more information or to make a proposal for the inclusion of a book. There is a simple contract that specifies the terms of use of the content. A sample contract can also be obtained by request at the above email address.

The complete contents of each book will be converted into XML according to the NCBI Book Document Type Definition (DTD), a public domain DTD developed at NCBI for this project (see below). Books may be submitted to conform to this DTD, or NCBI will convert the source data to validate against the Book DTD. If a conversion needs to be done, the content must be in a format robust enough to meet the needs of the BookShelf publishing system and DTD.

Any book that was printed from SGML or XML should allow for a straightforward conversion. We have had success converting books from Word, XYWrite, PDF, and Quark Express formats, and we anticipate that we would also be able to convert from other desktop publishing packages. HTML and PDF formats are less desirable because the data formats are less detailed.

Figures should be supplied in TIFF format, although GIF and JPEG formats may be accepted. The submitted text files are converted into XML according to the NCBI Book DTD; graphic files are converted into GIF and JPEG formats. Three hard copies of the book are also required, along with the electronic files.

The XML files are stored in a database. When a reader requests a book, chapter, or section, the XML is retrieved from the database and converted into HTML on the fly using Extensible Stylesheet Language Transformations (XSLT) and Cascading Style Sheets (CSS).

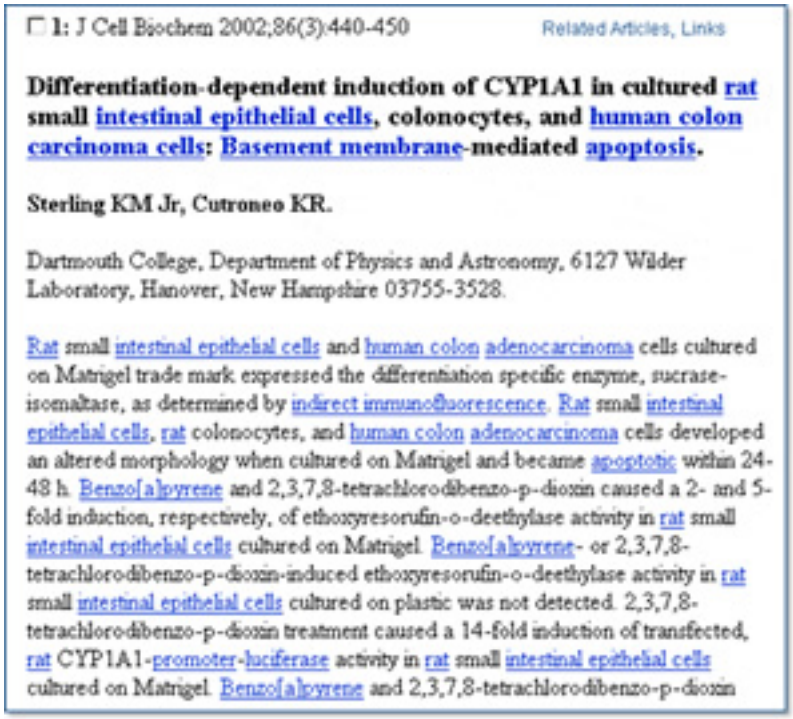
How to Use the Books

There are three ways to access the content in BookShelf:

1. Through hyperlinked terms in PubMed abstracts
2. By a direct search using search terms or phrases (in the same way as the bibliographic database of PubMed is searched)
3. Through the Table of Contents of the book (note: some publishers restrict browsing through the entire book by disabling hyperlinks in the Table of Contents)

Links from PubMed

The BookShelf can be accessed from all PubMed abstract pages. When viewing a full PubMed abstract, select the **Books** hyperlink in the upper right-hand corner. This generates a version of the abstract in which certain phrases and terms appear as hypertext links (see Figure 1). The linked term may be one or more words in length. If a word or a phrase is linked, it means that the exact phrase also appears in at least one book. Selecting a linked phrase retrieves a list of books that contain that term.


1: J Cell Biochem 2002;86(3):440-450 Related Articles, Links

Differentiation-dependent induction of CYP1A1 in cultured rat small intestinal epithelial cells, colonocytes, and human colon carcinoma cells: Basement membrane-mediated apoptosis.

Sterling KM Jr, Cutroneo KR.

Dartmouth College, Department of Physics and Astronomy, 6127 Wilder Laboratory, Hanover, New Hampshire 03755-3528.

Rat small intestinal epithelial cells and human colon adenocarcinoma cells cultured on Matrigel trade mark expressed the differentiation specific enzyme, sucrase-isomaltase, as determined by indirect immunofluorescence. Rat small intestinal epithelial cells, rat colonocytes, and human colon adenocarcinoma cells developed an altered morphology when cultured on Matrigel and became apoptotic within 24-48 h. Benzo[a]pyrene and 2,3,7,8-tetrachlorodibenzo-p-dioxin caused a 2- and 5-fold induction, respectively, of ethoxyresorufin-o-deethylase activity in rat small intestinal epithelial cells cultured on Matrigel. Benzo[a]pyrene- or 2,3,7,8-tetrachlorodibenzo-p-dioxin-induced ethoxyresorufin-o-deethylase activity in rat small intestinal epithelial cells cultured on plastic was not detected. 2,3,7,8-tetrachlorodibenzo-p-dioxin treatment caused a 14-fold induction of transfected, rat CYP1A1-promoter-luciferase activity in rat small intestinal epithelial cells cultured on Matrigel. Benzo[a]pyrene and 2,3,7,8-tetrachlorodibenzo-p-dioxin

Figure 1: A PubMed abstract showing terms linked to books.
This view was generated by selecting **Links** found in the *top right corner* of the abstract and then selecting **Books** from the drop-down menu that appears.

A statistical weighting system based on the frequency of each phrase in a book section, relative to the rest of the book, is used to identify "good" phrases. A phrase that appears repeatedly in only a few sections and rarely in other parts of the book indicates a definitive phrase for those few sections; therefore, it ranks highly. Furthermore, the appearance of a phrase in the title, for example, has a greater value in the weighting system than one appearing solely in the text.

Each PubMed abstract can thus be linked to the appropriate book pages. This method allows two very dissimilar types of text—the dense, focused PubMed abstracts and the more descriptive book text—to find common ground.

Direct Search of Books

Book contents may be searched directly from the BookShelf homepage by using the search boxes located in the Table of Contents and navigation bars of books or by selecting **Books** from the pull-down menu in any Entrez database search bar (see Chapter 14). Search terms may be combined using Boolean operators that conform to PubMed syntax (see Chapter 2). The BookShelf also allows search fields to be specified. A complete list of BookShelf database fields can be found in Table 1.

Table 1. Field limits for use in the BookShelf.

Field[a]	Use
[Author]	Search for the authors of books or chapters.

Field[a]	Use
[Book]	Typically used with a Boolean expression to limit a search to a particular book.
[PmId]	Locate a journal article citation in a book by its PubMed ID.
[Rid]	Locate a particular book element (such as a figure or table) by its reference ID.
[Secondary Text]	Search for secondary text, e.g., units (mg/l, etc.)
[Title]	Search for words used in any title (book, chapter, section, subsection, figure, etc.).
[Type]	Locate a division of a book such as a section, chapter, or figure group.

[a] Filters are applied immediately following a search term, with no separating spaces, e.g., watson[author] AND cmed[book].

Interpreting Results from a Search or PubMed Link

Results are shown as a list of books in which the term is found, along with the number of sections, figures, and tables that contain the term (Figure 2). The book that contains the most hits appears at the top of the list. Choosing the hyperlinked number of items that is associated with a particular book will then display a document summary list of the individual sections, figures, and tables found. (When a term or phrase is found fewer than 20 times within the BookShelf, the document summary page is shown directly, without first displaying the results clustered by book.)

Figure 2: A results page from a BookShelf search.
If more than 20 sections, tables, and/or figures are found that contain the query term, a summary page, such as the one above, is displayed.

The document summary list is sorted with the most relevant documents shown at the top of the page. The sorting makes use of scores allocated to phrases as a measure of how relevant they are to a given section (a part of the statistical weighting system also used for linking PubMed abstracts to the books). For each book section found, the title, along with some context regarding the hierarchy of the section location (e.g., the chapter and book), is given. An icon is used to distinguish figure legends and tables from text sections (Figure 3). Selecting a hyperlinked section title displays the part of the book that contains the search term. From this point, the user may be able to navigate further throughout the book content, according to the policy of the publisher (see *Navigating Book Content*).

Figure 3: A document summary list of book sections.
The most relevant sections to the query appear at the top of the list. Note the icons that designate figure and table hits. The list may also be displayed in a brief format that lists only the section names that contain the term by choosing **Brief** from the drop-down menu to the immediate *right* of the **Display** button.

Navigating Book Content

Each HTML page of content seen in a web browser represents one section of one chapter of a book, i.e., all of the content (including subsections and so on) within the first-level heading of a chapter. The amount of content this represents varies according to the structure of the original book. Some books have very long sections, some short, some a mixture; although on the whole, most chapters are divided into 3-10 sections.

The top of every page contains links to both short and detailed Tables of Contents and a description of the current location within the book (Figure 4). The hierarchal elements that describe the current location are hyperlinked and may be used to travel up the organizational levels of a book. Additionally, a navigation sidebar shows the current section among its peers and lists the figures and tables found within the current section

(Figure 4). Reference citations in the text are linked where possible to PubMed abstracts by the Citation Matcher. References internal to the book, e.g., to other chapters or sections, figures, tables, and boxes, are also hyperlinked. Further navigation from the current page to other parts of the book depends on the access policy of the publisher.

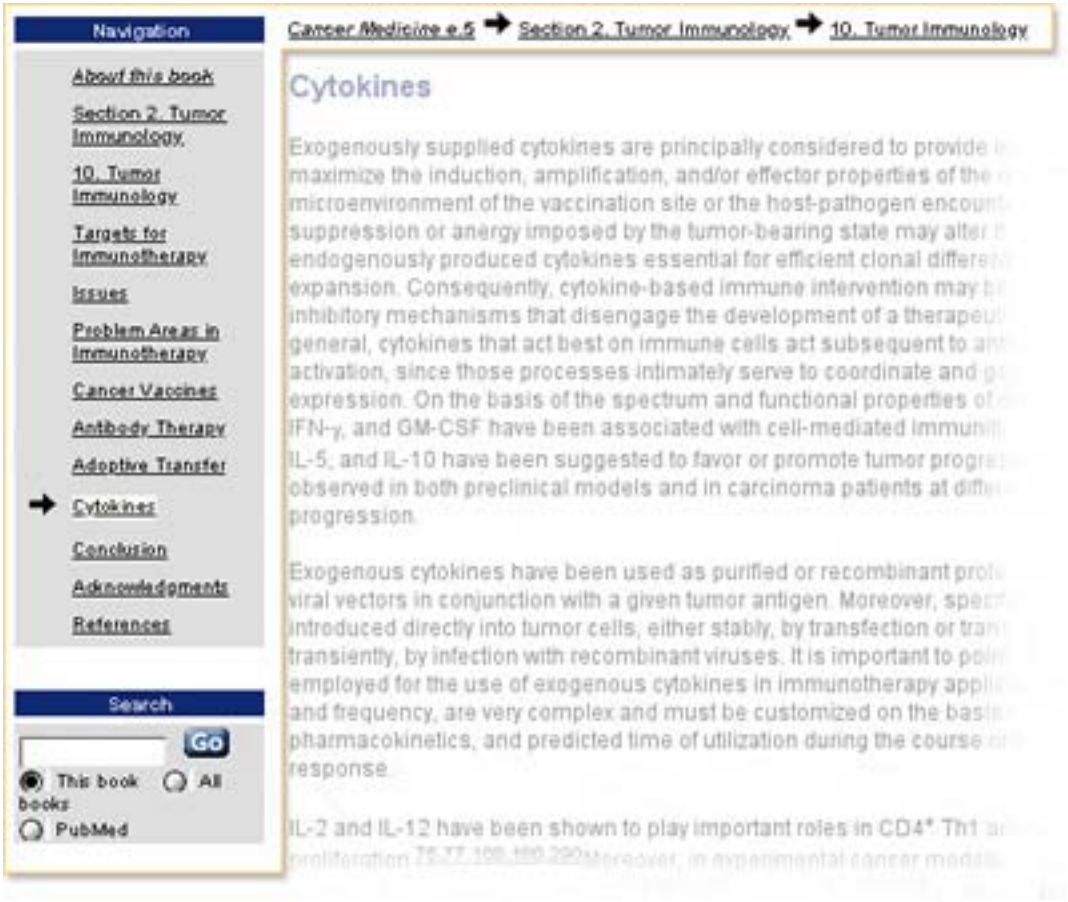


Figure 4: Navigation elements that appear on every book HTML page.
Hyperlinks to various sections within a chapter appear within a navagation bar to the *left* of each page. Hyperlinks may be disabled within some books at the request of the publisher. A search box is located *below* the navigation bar. At the *top* of each page is a hyperlinked, hierarchical tree that illustrates a page's relative position within a book.

Technology

Text Conversion and XML

All book content submitted to the BookShelf is converted to XML according to the public NCBI Book DTD.

Any files submitted in SGML or XML are converted to the Book DTD using XSLT. Books submitted in desktop publishing or word processing formats are converted by a contractor using proprietary technologies. Once the book XML is valid against the DTD, the book is ready to be loaded into the database.

The XML for a book is generally a set of files. Each chapter and appendix is an independent file, as is the frontmatter. The book is pulled together by the book.xml file, which defines the structure of the book. For example, a book with two chapters and a bibliography at the end of the book would be structured as follows:

```
<!DOCTYPE book SYSTEM "ncbi-book.dtd" [
<!-- graphics -->
<!ENTITY % Graphics SYSTEM "graphics.xml">
%Graphics;
<!ENTITY frontmatter SYSTEM "fm.xml">
<!ENTITY chapter1 SYSTEM "ch1.xml">
<!ENTITY chapter2 SYSTEM "ch2.xml">
<!--back-->
<!ENTITY biblist SYSTEM "biblist.xml">
]>
<book>
&frontmatter;
<body>
&chapter1;
&chapter2;
</body>
<back>
&biblist;
</back>
</book>
```

The book.xml file is composed of two parts. The first part defines all of the components that are required to build the book. These definitions occur within the <! DOCTYPE [ ] > tag. The second part builds the structure of the book. The root element is <book>. <book> contains whatever is in &frontmatter;, <body>, and <back>.

The book.xml example above refers to five external files: fm.xml, which contains all of the frontmatter of the book; ch1.xml, which contains chapter 1; ch2.xml, which contains chapter 2; biblist.xml, which contains the bibliography for the book; and graphics.xml, which defines the images. If any of these files is not valid according to the DTD or if the files are not found where they are defined in their <!ENTITY > declaration, then the book will not be valid.

Images

All of the images, including figures, icons, and book-specific character graphics (see Special Characters below), are called out in the text as entities. The entities are defined in the graphics.xml file.

<!ENTITY ch2fu6 SYSTEM "data/mga/pictures/ch2/ch2fu6.gif" NDATA GIF>
<!ENTITY ch2fu7 SYSTEM "data/mga/pictures/ch2/ch2fu7.jpg" NDATA JPG>
<!ENTITY ch2fu8 SYSTEM "data/mga/pictures/ch2/ch2fu8.gif" NDATA GIF>
<!ENTITY ch2fu9 SYSTEM "data/mga/pictures/ch2/ch2fu9.gif" NDATA GIF>
<!ENTITY ch2fu10 SYSTEM "data/mga/pictures/ch2/ch2fu10.gif" NDATA GIF>
<!ENTITY ch2e1 SYSTEM "data/mga/pictures/ch2/ch2e1.gif" NDATA GIF>
<!ENTITY ch2e2 SYSTEM "data/mga/pictures/ch2/ch2e2.gif" NDATA GIF>

Graphic files are converted into GIF and JPEG formats and optimized for display on the Web. The images are not loaded into the database; they are retrieved from a file server when called by the HTML page.

Math and Formulae

Math expressions and chemical formulae and structures are handled as images.

Special Characters

The BookShelf uses the same character sets that PubMed Central uses (see Chapter 9). These include a number of standard ISO character sets (8879 and 9573), along with a set of characters that has been defined to accommodate characters not in the standard set. The ISO Standard Character sets referenced are listed in Box 1 of Chapter 9. Special characters are converted to the BookShelf/PubMed Central (PMC) character set during conversion into XML. Characters created for one book (book-specific characters) are called out in the XML as images. To provide for the most flexibility in displaying characters across platforms, BookShelf uses UTF-8 encoding whenever possible. Because not all browsers support the same subset of UTF-8 characters and some characters cannot be represented in UTF-8, the BookShelf displays characters as a combination of GIFs and UTF-8 characters, depending on the Browser/OS combination and the character to be displayed.

The BookShelf Data Flow

The XML files for each book are stored in a Sybase database. To show the best representation to each reader, the reader's browser and operating system are noted and passed to the rendering software. When a reader requests a book, chapter, or section, the XML and the character images or UTF-8 characters appropriate to the reader's system are retrieved from the database.

The XML is converted to HTML using XSLT stylesheets. The look of the HTML pages is controlled further using CSS, which allow manipulation of colors, fonts, and typefaces (Figure 5).

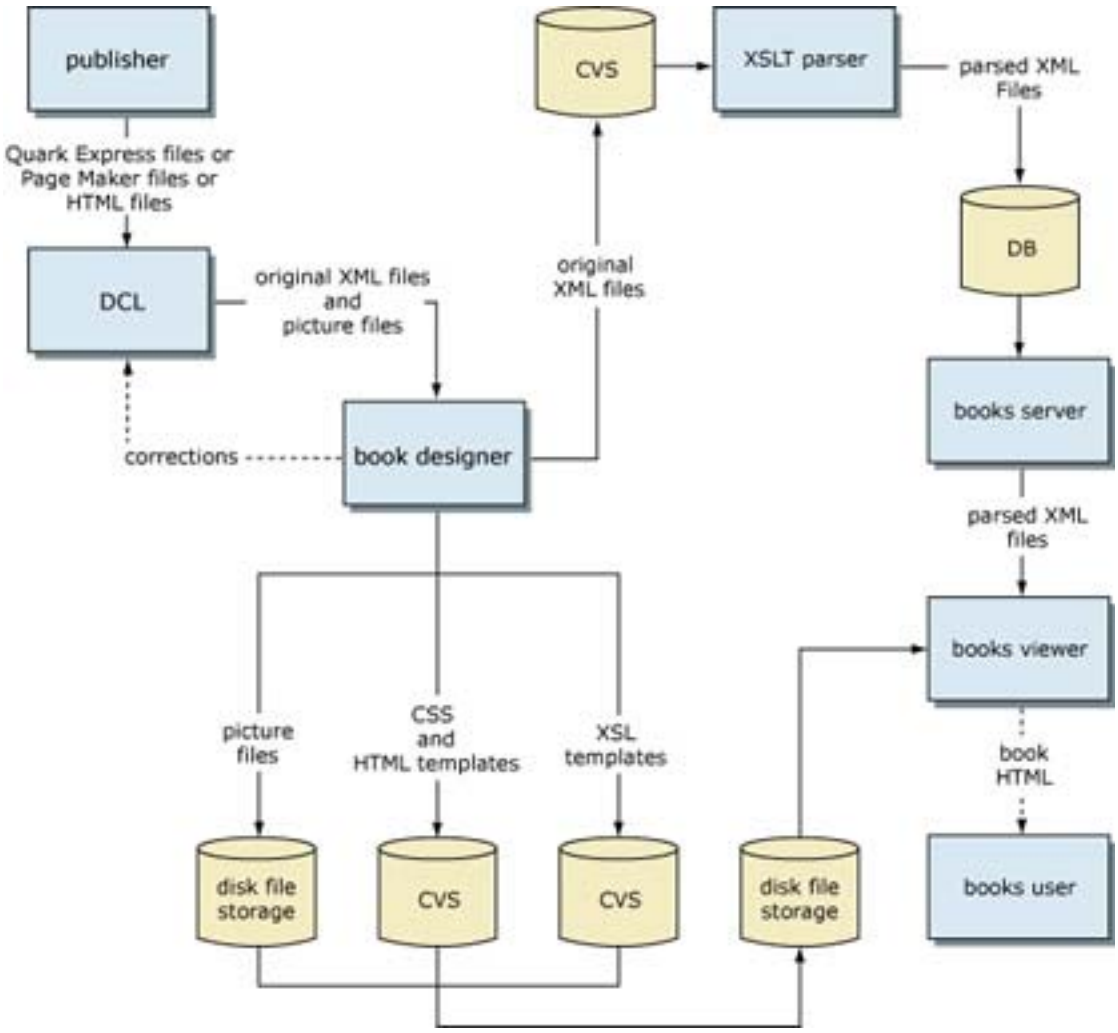


Figure 5: Processing and data flow of books.

The Table of Contents for a book is created from actual elements within the book content, rather than from the Table of Contents given in the book frontmatter of the hard copy. This ensures that the Table of Contents represents the content accurately as it is organized on BookShelf.

NCBI Book DTD

History

In the first version of the NCBI BookShelf project, Quark files were converted directly to HTML for display online. The result was effective, illustrating the value of having a textbook online and linked to PubMed; however, it was labor intensive, limiting, and not scaleable.

To simplify the delivery of books online and to allow for the expansion of linking within the Entrez system, NCBI decided to convert all content into a centralized XML format. The normalized XML content is easier to render, allows added value such as the addition of links to other NCBI databases, and simplifies the addition of new volumes.

PMC created a new DTD for the BookShelf project, which was based on the ISO-12803 DTD. As more books were converted to the NCBI Book DTD, changes had to be made to accommodate the data.

The NCBI Book DTD is a public DTD available on request from books@ncbi.nlm.nih.gov.

Frequently Asked Questions

1. How do I access the books at NCBI?

The online books can be accessed by direct searching in Entrez or through PubMed abstracts. After performing a general PubMed search, click on the author name of one of the search results to view the abstract. A hypertext link called **Links** is displayed to the right of the abstract title. This link contains a drop-down menu consisting of various choices, depending upon the specific abstract. Choosing **Books** from this drop-down menu will highlight keywords in the abstract that, when selected, initiate a search of all BookShelf content for that particular term.

2. Which books are available at NCBI?

The book list is updated on a regular basis and can be viewed on the BookShelf homepage: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Books.

3. Can I search the books at NCBI?

Yes, the books can be searched either as a complete collection or as a single, selected book (restricted using search options found under **Limits**).

4. Can I browse the whole book?

The system has been designed so that the user is delivered to the most relevant book sections for a particular term or concept. Although navigation is possible in the immediate vicinity of the page to which you are delivered, it may not be possible to browse the complete book on BookShelf. The range of navigation for each book is determined on a case-by-case basis, in agreement with the publisher.

5. I am the publisher/author/editor of a book. How can I participate?

Please email books@ncbi.nlm.nih.gov to discuss potential projects.